บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงรายละเอียดของทฤษฎีที่เกี่ยวข้องและเทคนิคต่างๆที่นำมาใช้ในงานวิจัยซึ่งมีรายละเอียดดังต่อไปนี้

2.1 ลักษณะของภาษาไทย

ภาษาไทยมีลักษณะแตกต่างจากภาษาอังกฤษ เนื่องจากในภาษาไทยมีการเขียนติดกันไปทั้งประโยค อีกทั้งคำในภาษาไทยคำหนึ่งอาจประกอบไปด้วยสระที่เป็นสระประกอบ คือมาจากสระอื่นอีกหลายตัวประกอบกัน เช่น สระเอียะ สระเอือะ เป็นต้น (ปโยธร อุราธรรมกุล กานดา รุณนะพงศา, 2549, น. 3)

ชื่อเฉพาะหรือวิสามานยนามตามความหมายจากพจนานุกรมฉบับราชบัณฑิตยสถาน (2542) หมายถึงคำนามที่เป็นชื่อเฉพาะตั้งขึ้นไว้สำหรับเรียกคน สัตว์ สิ่งของ และสถานที่เพื่อให้รู้ ชัดว่าเป็นใครหรืออะไร เช่น นายสมชาย (ชื่อคน) เป็นต้น ชื่อเฉพาะนี้อาจมีคำเรียกเป็นภาษาอังกฤษได้หลายแบบ ได้แก่ Proper Name, Named entity และ Specific Name เป็นต้น ซึ่งแต่ละแบบอาจใช้เป็นคำเรียกโดยคนต่างกลุ่มกันและอาจต้องการเน้นย้ำสิ่งที่ต่างกัน เช่น นักวิจัยทางด้านภาษาศาสตร์คอมพิวเตอร์มักจะใช้คำว่านิพจน์ระบุนาม (Named Entity) ในขณะที่นักภาษาศาสตร์มักจะใช้คำว่าชื่อเฉพาะ (Proper Name) แต่โดยพื้นฐานแล้วคำเรียกภาษาอังกฤษเหล่านี้มีลักษณะเหมือนกัน (สุฤดี ฉัตรไตรมงคล,2548, น.7)

นิพจน์ระบุนามเป็นหน่วยภาษาที่จัดว่าเป็นหน่วยหลักหน่วยหนึ่งที่ทำให้การสื่อสารของมนุษย์มีประสิทธิภาพและนิพจน์ระบุนามเป็นหน่วยภาษาที่มีคุณสมบัติเฉพาะตัวที่มีความสำคัญต่อการประมวลผลภาษาธรรมชาติ ซึ่งคุณสมบัติอันหนึ่งที่เป็นคุณสมบัติร่วมในทุกภาษาคือ นิพจน์ระบุนาม เกิดขึ้นได้เสมอและเกิดได้กับทุกสิ่งทุกอย่าง และไม่มีตัวบ่งชี้สำหรับระบุแสดงว่ารูปภาษานั้นเป็นนิพจน์ระบุนามเหมือนที่มีการกำหนดไว้ในบางภาษา เช่น ภาษาอังกฤษ ที่ใช้การขึ้นต้นด้วยตัวอักษรพิมพ์ใหญ่เป็นตัวบ่งชี้ นิพจน์ระบุนามในภาษาไทยสามารถสร้างขึ้นจากคำที่มีใช้อยู่แล้วในภาษา เช่น "สอน" นำไปสร้างเป็นนิพจน์ระบุนามเป็น"นายสอน" "ต้นพะยอม" นำไปสร้างเป็นนิพจน์ระบุนาม "นางสาวพะยอม"ซึ่งการสร้างนิพจน์ระบุนามขึ้นมานั้นจะใช้คำจำนวนเท่าใดก็ได้ไม่มีการกำหนดกฎเกณฑ์ไว้ (อมรทิพย์ กวินปณิธาน,2546, น.9-10) นิพจน์ระบุ

นามเป็นปัญหาสำคัญในการประมวลผลภาษา เพราะนิพจน์ระบุนามนั้นมีได้ไม่จำกัด จึงไม่สามารถจัดเก็บในพจนานุกรมได้ครบถ้วน และนิพจน์ระบุนามอาจประกอบด้วยคำทั่วไปทำให้เป็นปัญหาในการระบุขอบเขตของนิพจน์ระบุนาม การไม่สามารถหานิพจน์ระบุนามได้ถูกต้องนี้ จะส่งผลให้เกิดข้อผิดพลาดในการประมวลผล การตัดคำภาษาไทย (Thai Word Segmentation)

2.2 วิธีที่ใช้ในการตัดคำ

วิธีที่ใช้ตัดคำแบ่งออกเป็น 3 ประเภทใหญ่ๆ ได้แก่ การใช้กฎ การใช้พจนานุกรม และการใช้คลังข้อความ (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 4)

2.2.1 การใช้กฎ

การตัดคำโดยการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะการประสมอักษร ลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกต้องของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำ แต่ข้อดีของวิธีนี้คือมีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2.2.2 การใช้พจนานุกรม

การตัดคำโดยพจนานุกรมเป็นการตัดคำโดยใช้สายอักขระ (String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้จะต้องทำการจัดเก็บคำไว้ในพจนานุกรม วิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลามากกว่า

2.2.3 การใช้คลังข้อความ

การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้ คลังข้อมูลแบ่งออกเป็น 2 วิธี คือการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic Word Segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-based Word

Segmentation) วิธีการตัดคำในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดย วิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุด

ตารางที่ 2.1

ตารางเปรียบเทียบการตัดคำแต่ละวิธี

วิธีการ	ข้อดี	ข้อเสีย
การใช้กฎ	- มีความรวดเร็ว ในการตัดคำ - ไม่ต้องการข้อมูลในหน่วยความจำของคอมพิวเตอร์	- ยากที่จะสร้างกฎไวยากรณ์ได้สมบูรณ์ - ทำได้ในระดับพยางค์ - ไม่สามารถจัดการคำที่เป็น คำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words)
การใช้พจนานุกรม	- มีความแม่นยำสูงในการตัดคำ - สามารถแก้ไขปัญหา คำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words) โดยการเพิ่มเข้าไปในพจนานุกรม	- ต้องการหน่วยความจำมากเพื่อจัดการกับการเก็บพจนานุกรม - บริหารจัดการยากที่จะนำคำ ที่เป็นคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words) เข้าไปในพจนานุกรม
การใช้คลังข้อความ	- ไม่จำเป็นต้องเก็บพจนานุกรมในหน่วยความจำ - แก้ไขปัญหาคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words)ได้บางกรณีโดยการเรียนรู้จากคลังข้อความ - สร้างกฎไวยากรณ์ได้จากการเรียนรู้จากคลังข้อความ	- ต้องการคลังข้อความที่มีขนาดใหญ่และเวลาในการเรียนรู้และต้องการอัลกอริทึม (Algorithm) ที่ดีในการเรียนรู้

2.3 เทคนิคการตัดคำ

เทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปมี 5 วิธี ซึ่งมีวิธีการใช้นำไปใช้แตกต่างกันขึ้นอยู่กับลักษณะของคำเพื่อนำไปใช้การตัดคำให้มีประสิทธิภาพที่ดีที่สุด

2.3.1 วิธีการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching)

วิธีนี้จะทำการตรวจสอบสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม จะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อยๆ จนจบสายอักขระตัวอย่างคำว่า “กองกลาง” การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ก , กอ และคำว่า กอง ส่วนคำว่า กองก ไม่พบอยู่ในพจนานุกรม ดังนั้นจึงได้คำว่า กอง ซึ่งเป็น คำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ กลาง เมื่อนำไปค้นในพจนานุกรมจะได้ว่า ก , กล , กลาง ดังนั้นจึงเลือกคำว่า กลาง คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น กอง|กลาง (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 5)

2.3.2 วิธีการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching)

วิธีการนี้คล้ายกับวิธีการเทียบคำที่ยาวที่สุด เพียงแต่จะเลือกคำที่สั้นที่สุดที่พบก่อน แต่วิธีนี้พบว่าได้จำนวนคำมากที่สุดแต่ความถูกต้องของคำหลังทำการตัดคำแล้วน้อยกว่าการใช้วิธีเทียบคำที่ยาวที่สุด ตัวอย่างคำว่า “โคลงเรือ” การตัดคำโดยวิธีนี้จะเลือกเอาคำแรกที่ค้นหาเจอพจนานุกรม ดังนั้นจะได้ว่า โค ลง เรือ (โดยไม่เลือกคำว่า “โคลง” ที่จะพบต่อไปภายหลังหากทำการค้นหาต่อ) วิธีนี้ใช้เวลาน้อยกว่าการเทียบคำยาวที่สุด แต่ความถูกต้องที่ได้การตัดคำแบบเทียบคำยาวที่สุดจะมากกว่า (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 5)

2.3.3 วิธีการตัดคำที่ใช้ความถี่ของคำ หรือ สถิติ (Probabilistic Word Segmentation)

วิธีการนี้เป็นแนวทางหนึ่งในการแก้ปัญหาคำกำกวมของประโยคภาษาไทยโดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวันโดยจัดเรียงคำในพจนานุกรมตามความถี่ที่พบ และ

ใช้วิธีการตัดคำแบบเดียวกับข้อ 2.3.1 และ 2.3.2 ตัวอย่างคำว่า “ก็อด” ในกรณีนี้หากใช้ความถี่ของคำจะได้ว่า ก็ อด มีความถี่สูงกว่า ก็อด (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา,2549,น. 5)

2.3.4 **วิธีการย้อนรอยกลับ** (Back Tracking)

เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำกับคำที่มีอยู่ในพจนานุกรม อาจพบกรณีที่คำที่พบมีมากกว่า 1 คำแล้วทำการเลือกคำที่ยาวที่สุดทำให้สายอักขระที่ตามมาจากคำนั้นไม่สามารถตัดคำได้ เนื่องจากไม่พบตามพจนานุกรม กรณีนี้จะทำการย้อนไปอีกคำที่ไม่ถูกเลือกแล้วทำการตัดคำต่อไปตัวอย่างเช่นคำว่า “เมื่อยามนี้” การเปรียบเทียบกับพจนานุกรมจะได้ว่า เมื่อ , เมื่อย ดังนั้นจึงเลือกคำที่ยาวที่สุดจะได้คำว่า เมื่อย ส่วนที่เหลือคือ –ามนี้ ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้นจะทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่งคือ เมื่อ จะได้เป็น เมื่อ ยาม นี้ (โดยคำว่า ยามเกิดจากการเลือกคำที่ยาวที่สุดระหว่าง ยา และ ยาม) (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 5)

2.3.5 **วิธีการตัดคำแบบใช้คุณลักษณะ** (Feature - based Approach)

วิธีการนี้จะพิจารณาจากบริบท (Context Words) และการเกิดร่วมกันของคำ หรือหน้าที่ของคำ (Collocation) เข้ามาช่วยในการตัดคำตัวอย่างเช่น"ตากลม" ถ้าพบคำว่า "แป๋ว" ในบริบทก็จะสามารถตัดคำได้ว่า "ตา" "กลม" วิธีการนี้จำเป็นที่จะต้องมีฐานข้อมูลเป็นจำนวนมาก และจะต้องมีการเรียนรู้การสร้างคำในบริบท หรือการเกิดร่วมกันของคำแต่ละคำเพื่อให้มีข้อมูลที่จะนำมาใช้ในการตัดคำ ซึ่งจากวิธีการข้างต้น เรายังพบว่าแต่ละวิธีการมีข้อจำกัดและไม่สามารถแก้ปัญหาได้ทั้งหมด ในปัจจุบันจึงยังมีการพัฒนาและคิดค้นวิธีการใหม่ๆ ที่จะช่วยในการแบ่งคำให้มีประสิทธิภาพมากยิ่งขึ้น

2.4 **งานวิจัยที่เกี่ยวข้อง**

2.4.1 **ยุคการใช้กฎ**

ในยุคการใช้กฎคอมพิวเตอร์ยังไม่สามารถประมวลผลได้สูงมากนัก ประกอบกับหน่วยความจำในคอมพิวเตอร์มีขนาดเล็ก ทำให้เกิดการพัฒนาการใช้กฎในการตัดพยางค์ขึ้นมา

ก่อน เนื่องจากพยางค์นั้นมีกฎเกณฑ์ที่แน่นอนมากกว่าคำ ทำให้ในยุคนี้มีการนำกฎเข้ามาใช้ในการตัดพยางค์ (ไพศาล เจริญพรสวัสดิ์,2541,น. 5)

1.งานของสุรินทร์ จรรยาพรพงษ์ (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 6)

เป็นงานวิจัยที่ตัดคำโดยใช้กฎ โดยกฎที่นำมาใช้นั้นได้มาจากไวยากรณ์ภาษาไทยและวิเคราะห์ ลักษณะของการใช้ โดยลักษณะของกฎแบ่งได้เป็น 2 ชนิด คือ กฎการหาขอบเขตหน้า (Front Boundary Recognition Rule) และกฎการหาขอบเขตหลัง (Tail Boundary Recognition Rule) และในแต่ละกลุ่มยังแบ่งออกเป็น 2 กลุ่มย่อยๆ คือแบ่งตามคุณสมบัติของตัวอักษรโดยกฎที่ได้ออกมานี้จะจัดให้อยู่ในกลุ่มเอ (Group A) และแบ่งตามคุณสมบัติของรูปแบบการใช้สระแต่ละตัวซึ่งกฎที่ได้จะจัดไว้ในกลุ่มบี (Group B)

1.1 กฎที่ได้จากคุณสมบัติของอักษรในการหาขอบเขตหน้าของพยางค์

กฎ A-1F : สระต่างๆ เหล่านี้ อะ อา อิ อี อึ อึ อุ อู ไต่คู้ อำ ไม้หันอากาศ และวรรณยุกต์ทั้งหมด จะต้องมีพยัญชนะอยู่ข้างหน้าอย่างน้อย 1 ตัวอักษรเสมอ

กฎ A-2F : สระ เ แ ไ โ จะเป็นตัวอักษรแรกของพยางค์เสมอยกเว้นมีพยัญชนะนำหน้า

กฎ A-3F : สระ ใ จะเป็นตัวอักษรแรกของพยางค์เสมอ โดยไม่มีข้อยกเว้น

กฎ A-4F : พยัญชนะ ฉ ผ ฝ ฮ จะต้องเป็นพยัญชนะต้นเสมอยกเว้นมีสระเหล่านี้นำหน้า เ แ ไ โ ใ นำหน้า

กฎ A-5F : มีเพียง 9 คำในภาษาไทยเท่านั้นที่ใช้ ฤ เป็นพยัญชนะต้น เช่น ฤทัย ฤดี นอกนั้นจะพบ ฤ เดี่ยวๆ ยกเว้นคำว่า อมฤต

กฎ A-6F : ห มักพบเป็นคำนำหน้าพยางค์ ยกเว้นบางคำเช่น สห มหา คหบดี มหกรรม มรสพ คหกรรม เป็นต้น

กฎ B-1F : จากกฎ A1-F ถ้ามีคำที่นำด้วยพยัญชนะสองตัว จะทำการตรวจสอบ ว่าบางตัวอาจทำหน้าที่เป็นสระ

1.2 กฎที่ได้จากคุณสมบัติของอักษรในการหาขอบเขตหลังของพยางค์

กฎ A-1T : พยัญชนะต่อไปนี้ ศ ณ ญ ษ ฐ ฎ ฏ ฒ ฬ ฌ จะเป็นตัวสะกดเสมอ ยกเว้นพยางค์ต่อไปนี้ ศก ศร ญวน ศตวรรษ และพยางค์อื่นๆ อีกแต่สามารถจัดการโดย A-1F

กฎ A-2T : สระ อึ จะต้องมีตัวสะกดหนึ่งตัวเสมอ ยกเว้น ฤ ฺหึ ฮึ

กฎ A-3T : ไม้หันอากาศ จะต้องมีตัวสะกดอย่างน้อยหนึ่งตัวเสมอ

กฎ A-4T : สระ อ และสระ อำ จะต้องตามพยัญชนะเสมอ

กฎ A-5T : การันต์ มักจะปรากฏเป็นตัวสุดท้ายของพยางค์ยกเว้นบางคำเช่น ปาล์ม ฟาร์ม

กฎ A-6T : ไม้หันอากาศจะต้องปรากฏระหว่างพยัญชนะสองตัว เช่น คัน ขัน

1.3 ตัวอย่างกฎที่ได้จากคุณสมบัติการใช้สระในการหาขอบเขตหลังของพยางค์

กฎ B-1T : สระเหล่านี้ ไม้หันอากาศ อิ อึ อี ถ้ามีวรรณยุกต์แล้วจะต้องมีตัวสะกด 1 ตัว เสมอ กฎ B-2T : สระ อี ที่ปรากฏวรรณยุกต์นอกจากไม้ตรี ส่วนใหญ่จะไม่มีตัวสะกด

กฎ B-3T : สระ ็ เ- ็- แ ็ -็- อ- และ เ ิ - ต้องมีตัวสะกด 1 ตัว

กฎ B-4T : สระในรูปแบบดังต่อไปนี้ อัว อัย อัวะ อือ เอา เอาะ เอะ แอะ โอะ เอียะ เออะ ไม่ต้องการตัวสะกด ยกเว้นบางพยางค์ในรูปแบบดังนี้ เอา- เออ- โดยเครื่องหมาย – แทนพยัญชนะ 1 ตัวอักษร

กฎ B-5T : มีเพียง 20 คำที่ใช้ ใ

กฎ B-6T : สระจากคำ ไ- ที่มีวรรณยุกต์ จะครบพยางค์ เนื่องจาก ไ ไม่มี ตัวสะกด มีเพียงบางคำที่นอกเหนือจากนี้ เช่น ไฮ่ย์

งานวิจัยนี้ยังสามารถตัดคำให้อยู่ในรูปแบบพยางค์หรือคำสั้นๆ เพียงแค่นั้น นอกจากนี้กฎที่มีอยู่ยังไม่สามารถรองรับคำที่มีการสะกดแตกต่างออกไป หรือ นิพจน์ระบุนามเป็นต้น

2.4.2 การใช้พจนานุกรม

เนื่องจากในยุคนี้เครื่องคอมพิวเตอร์ได้พัฒนาขึ้นและมีหน่วยความจำมากขึ้น จากยุคที่แล้วนำเอากฎเข้ามาช่วยแบ่งพยางค์ แต่สำหรับการแบ่งคำการใช้กฎอย่างเดียวไม่สามารถหาขอบเขตของคำได้ ทำให้มีการคิดค้นหาวิธีการแบ่งคำโดยมีการเอาพจนานุกรมมาใช้ร่วมกับกฎในการตัดคำด้วย (ไพศาล เจริญพรสวัสดิ์,2541,น. 8)

1. งานของยืน ภู่วรวรรณและวิวรรธน์ อิ่มอารมณ์ (ปโยธร อุราธรรมกุล,กานดา รุณนะพงศา,2549, น. 8)

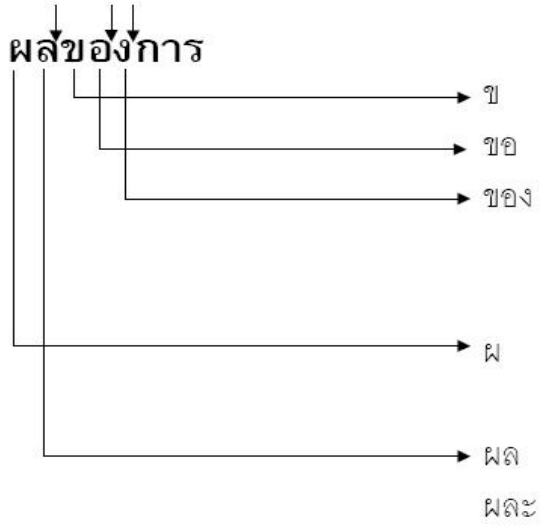
ในงานวิจัยนี้จะเป็นงานวิจัยการแบ่งพยางค์ด้วยพจนานุกรม ซึ่งถือได้ว่าเป็นงานวิจัยแรกของการตัดพยางค์ที่มีการนำพจนานุกรมเข้ามาใช้ โดยจะเก็บพยางค์ต่างๆไว้ในพจนานุกรม และมีการนำกฎไวยากรณ์ต่างๆจำนวน 18 กฎเข้ามาช่วยในกรณีที่ไม่พบพยางค์ในพจนานุกรม

หลักการทำงานของกระบวนวิธีการตัดพยางค์ด้วยพจนานุกรมนี้ก็คือ จะทำการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่ได้เก็บไว้ในพจนานุกรม ใน

กรณีที่ทำการตรวจสอบแล้วพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้ทำการเลือกแบ่งพยางค์โดยเลือกพยางค์ยาวที่สุด แล้วค่อยทำไปเรื่อยจนจบสายอักขระ แต่ในกรณีที่เลือกพยางค์ที่ยาวที่สุดไปแล้ว ทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้ย้อมกลับ (Back Tracking) ไปเลือกพยางค์ที่ยาวรองลงมาแทน ซึ่งวิธีการนี้จะเป็นที่รู้จักในชื่อการตัดคำแบบเลือกพยางค์ยาวที่สุด (Longest Matching)

ภาพ 2.1

ภาพแสดงตัวอย่าง การตัดคำ แบบ Longest Matching



งานวิจัยนี้ได้นำพจนานุกรมร่วมกับการใช้กฎเพียง 18 กฎซึ่งไม่เพียงพอต่อการตัดคำที่เป็นนิพจน์ระบุนามหรือคำที่เกิดขึ้นใหม่ งานวิจัยนี้จะใช้ได้กับคำที่อยู่ในพจนานุกรมเพียงอย่างเดียว

2. งานของ กานดา รุณนะพงศา และ ปโยธร อุราธรรมกุล

หลักการทำงานจะใช้กฎในการตัด โดยมีกฎ 2 ส่วน คือ กฎที่ใช้ในการตัดพยางค์ มีอยู่ 6 กฎ และ กฎ สำหรับตัดคำภาษาไทย มีอยู่ 47 กฎ และมีการจัดเก็บพจนานุกรมแบบใหม่

ตารางที่ 2.2

ตัวอย่างคำศัพท์ในพจนานุกรมแบบใหม่

คำ	การจัดเก็บ
กระ	*[กระ]*
กระจายเสียง	[กระ]จาย[เสียง]
กระวนกระวาย	[กระ] [วน] [กระ]วาย
กระเบียดกระเสียร	[กระ] [เบียด] [กระ]เสียร
เบียด	*[เบียด]*
เบียดเสียดเยียดยัด	[เบียด] [เสียด] [เยียด] [ยัด]
ยัดเยียด	[เยียด] [ยัด]
ตกกระ	[ตก] [กระ]

จากตารางที่ 2.2. ในส่วนที่อยู่ใน [ ] คือรหัสของคำที่มีอยู่แล้วในพจนานุกรม ส่วนเครื่องหมาย * หมายถึงมีคำที่ใช้คำนี้อยู่ใน พจนานุกรมอีก เช่นคำว่า กระ มีการนำไปใช้ เช่น กระ อาจมีคำต่อท้าย เป็นกระจายเสียง ได้ ตกกระ อาจมีคำต่อท้ายเช่น ตกกระป๋อง

งานวิจัยนี้แบ่งประเภทเอกสารที่นำมาทำการทดลองตัดคำภาษาไทยออกเป็น 5 กลุ่ม ซึ่งมีความหลากหลายด้านเนื้อหาซึ่งเอกสารแต่ละประเภทจะประกอบไปด้วยทั้งภาษาไทยและภาษาต่างประเทศ คือ ข่าวเศรษฐกิจ ข่าวต่างประเทศและกีฬา ข่าวอื่นๆ บทความวิชาการ บทความวารสารทั่วไป ผลที่ได้จากการตัดคำสูงสุดคือ บทความวารสารทั่วไป มีค่า F1 เท่ากับ 93.75% การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่มุ่งเน้นเพื่อแก้ไขปัญหาการตัดคำด้วย กฎที่มีอยู่เดิมซึ่งไม่ครอบคลุมคำในภาษาไทยซึ่งมีคำที่มีลักษณะที่แตกต่างซับซ้อนออกไปจากเดิม บทความทางวิชาการซึ่งจะมีคำที่มีความยาวหลายพยางค์ ทำให้การตัดคำด้วยกฎนั้นตัดได้ไม่ถูกต้อง

2.4.3 **ยุคการใช้คลังข้อความ**

ในยุคปัจจุบัน เครื่องคอมพิวเตอร์มีประสิทธิภาพมากขึ้น มีหน่วยความจำที่มากขึ้น จึงได้มีการพัฒนาการนำคลังข้อความ (Corpus) เข้ามาใช้ในการตัดคำในยุคนี้ได้มีการพัฒนาการ

ตัดคำวิธีใหม่ โดยนำความรู้ต่างๆ จากคลังข้อความมาประยุกต์ใช้ ด้วยนำวิธีการทางสถิติเข้ามาใช้ (ไพศาล เจริญพรสวัสดิ์,2541,น. 11) เช่น N-gram Conditional random fields (CRFs) และไตรแกรม (Tri-gram)

2.4.3.1 หลักการตัดคำ โดย N-Gram

N-Gram คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (Character Sequence)ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (Word Sequence)ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้ ซึ่ง N-Gram ได้ใช้หลักการของสถิติในหลายๆ ด้านมาประยุกต์ใช้ หน่วยที่ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียง คำ หรือ อักขระก็ได้และ Gram มีได้หลายขนาด แล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง N ในแบบจำลองเอ็นแกรมนี้ใช้ความยาวของชุดอักขระและคำที่เขียนเรียงกันแตกต่างกัน ได้แก่ 2-Gram, 3-Gram, 4-Gram ฯลฯ ถ้าจะประมาณค่าความน่าจะเป็นของชุดคำหรือชุดอักขระจากคลังข้อมูลโดยการใช้วิธี N-Gram ผลที่ได้มีดังนี้

N-Gram ถูกนำไปประยุกต์ใช้กับงานด้านการประมวลผลภาษาธรรมชาติ (NLP : Natural Language Processing) ซึ่งส่วนใหญ่จะนำไปใช้แก้ไขข้อจำกัดการตัดคำในภาษาไทย เนื่องจากในระบบการตัดคำด้วยพจนานุกรม ซึ่งมักจะพบปัญหาว่ากรณีที่คำที่ปรากฏในเอกสารนั้นไม่มีอยู่ในพจนานุกรมนั้น ในขณะที่วิธี N-Gram คือ การนำบางส่วนของข้อความนั้นออกมาเป็นหน่วยคำ (Term) ตามค่า N เพื่อใช้แทนการตัดคำ โดยทำให้ลดเวลาในการค้นหาคำในเอกสารกับคำในพจนานุกรม แต่ในภาษาไทยนั้นจะไม่สามารถกำหนดได้ว่า 1 ตัวอักษรคือ 1 Gram เนื่องจากภาษาไทยมีสระและวรรณยุกต์ ดังนั้นในภาษาไทยจึงถือว่ากรณีที่ตัวอักขระเป็นตำแหน่งที่มีสระและวรรณยุกต์อยู่ด้วยจะถือว่าเป็น 1 Gram โดยทั่วไปในภาษาไทยนิยมใช้การตัดคำแบบ 2, 3 และ 4 Gram

1.งานของวิจัยของ อัศนีย์ ก่อตระกูลและคณะ (ปโยธร อุราธรรมกุล, กานดา รุณนะพงศา, 2549, น. 13)

งานวิจัย Automatic Thai Unknown Word Recognition จากการวิเคราะห์หน่วยคำ (Morphological Analysis) สำหรับภาษาไทยนั้นจะพบปัญหาต่างๆ ดังนี้ ปัญหาคำกำกวมและปัญหาการสะกดคำผิดโดยปัญหาเหล่านี้จะทำให้ผลการตัดคำออกมาโดยไม่ถูกต้อง ซึ่งงานวิจัยนี้จะทำการลดผลลัพธ์ที่ไม่เหมาะสมออกไป

โดยงานวิจัยนี้จะสถิติเข้ามาแก้ไขปัญหาการตัดคำและการกำหนดหน้าที่ของคำโดยการนำโมเดล ไตรแกรม (Tri-Gram) เข้ามาช่วยในการแก้ไขปัญหาการตัดคำ

ภาพที่ 2.2

ภาพแสดงตัวอย่าง สูตร Tri-Gram ในการทดลอง

$$P(w) = \prod_{i=1}^{n} P(Wi, n)$$

$$= \prod P(Wi \mid W_{i-1}, W_{i-2}) \qquad (1)$$

งานวิจัยนี้จะเน้นที่เป็นเกี่ยวกับการวิเคราะห์ คำผิดและคำกำกวมและการนำวิธีการ n-gram เข้ามาแก้ไขปัญหาได้ ซึ่งอาจนำมาช่วยแก้ไขในการตัดคำกำกวมได้

2.4.3.2 Conditional Random Fields (CRFs) (พัชรินทร์ ทองอารีย์,2548,น.8)

Conditional Random Fields เป็นวิธีการเรียนรู้จาก คลังข้อความ (Corpus) และจะทำการแปลงข้อมูลนำเข้า (Input Text) ให้เป็น Feature Vector จะสร้างโหนด (Node) ที่ควรจะเป็นไปได้ทั้งหมด จากนั้นจะทำการเลือกโหนดที่น่าจะเป็นไปได้มากที่สุด มาเป็นคำตอบโดยข้อมูลจะทำการเรียนรู้จากเครื่อง (Machine Learning) จะมีการกำหนดประเภทของเป็น Type รายละเอียด Label เพื่อทำให้คอมพิวเตอร์ทำการเรียนรู้

ภาพที่ 2.3

ภาพแสดงตัวอย่าง รายละเอียดของแต่ละประเภทใน CRFs

Tag	Type	Example
c	Character that can be the final consonant in a word	กขฃคฆงจชซญฎฏฐฑฒฌณดตถท ธนบปพฟภมยรลวศษสฬอ
n	Character that cannot be the final consonant in a word	ฅ ฉ ผ ฝ ฌ ห ฮ
v	Vowel that cannot begin a word	ะ ◌ั ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู ◌ั ◌า ◌ำ ◌ๅ
w	Vowel that can begin a word	เ แ โ ใ ไ
t	Tonal character	◌่ ◌้ ◌๊ ◌๋
s	Symbol	ฯ ๆ ฯ.
d	Digit character	0-9
q	Quote character	'-' "-"
p	Space character inside a word	_
o	Other character	A-z

จากภาพที่ 2.3 จะเห็นได้ว่ามีจะกำหนดประเภทของแต่ละ อักขระซึ่งแบ่งออกได้เป็น 10 ประเภท ตัวอย่างเช่น ตัวเลข จะอยู่ในประเภท d (Digit Character) เป็นต้น

ภาพที่ 2.4

ภาพแสดงตัวอย่างการเตรียมข้อมูลให้ CRFs

ท	c	B
◌่	t	I
า	v	I
น	c	I
ผ	n	B
◌ิ	v	I
◌์	t	I
ฟ	c	I
◌้	v	I
ง	c	I
ด	c	B
ะ	v	I

จากภาพ 2.3 จะประกอบด้วย 3 คอลัมน์ โดยคอลัมน์ที่ 1 คือคำที่ต้องการให้คอมพิวเตอร์เรียนรู้ คอลัมน์ที่ 2 คือการกำหนดแต่ละประเภทให้กับอักขระโดยกำหนดรายละเอียดของแต่ละประเภท ตามภาพด้านบนที่กำหนดไว้ ส่วน คอลัมน์ที่ 3 คือการกำหนดในส่วน คอลัมน์ที่ 3 คือ การกำหนด Label ให้กับ อักขระ B คือการกำหนดว่าเป็นอักษรตัวแรกสุดของคำ สำหรับ I คือการกำหนดว่าอักขระนี้เป็นอักขระที่อยู่ในคำ

1.งานวิจัยของ John Lafferty, Andrew McCallum และ Fernando Pereira (พัชรินทร์ ทองอารีย์, 2548,น.21)

เป็นงานวิจัยที่นำเสนอเอาเทคนิค Conditional Random Fields (CRFs) มาช่วยในการแบ่งแยกข้อมูลที่อยู่ต่อเนื่องกัน โดยทำการเปรียบเทียบประสิทธิภาพกับวิธี Maximum Entropy Markov Models (MEMMs) โดยข้อมูลที่นำมาทำการวิจัยถูกสร้างจาก Hidden Markov Models (HMMs) ซึ่งมี 2,000 ข้อมูลที่ใช้ในการเรียนรู้ (Train) และ 500 ข้อมูลที่ใช้ในการทดสอบ ผลจากงานวิจัยนี้คือ CRFs มีข้อผิดพลาด (error) น้อยกว่า MEMMs โดย CRFs มีค่าผิดพลาดเท่ากับ 4.6% ในขณะที่ MEMMs มีค่าผิดพลาดถึง 42%

2.งานวิจัยของ Fuchun Peng, Fangrfang Feng และ Andrew McCallum (2004) (พัชรินทร์ ทองอารีย์. 2548 น.21)

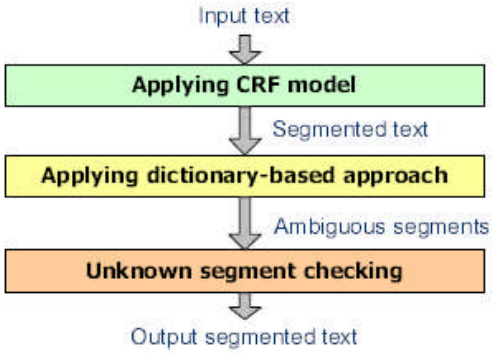
การตัดคำในภาษาจีนก็ประสบปัญหาเดียวกับการตัดคำในภาษาไทย เพราะลักษณะของภาษาจีนนั้นเป็นภาษาที่เขียนติดกันไปทั้งประโยคเช่นเดียวกับภาษาไทย จึงมีงานวิจัยที่ศึกษาเกี่ยวกับวิธีการต่างๆ ในการตัดคำของภาษาจีน สำหรับในงานของ Fuchun Peng, Fangfand Feng และ Andrew McCallum นั้น ได้ทำการสาธิตเอาเทคนิค Conditional Random Fields (CRFs) มาประยุกต์ใช้ในการตัดคำของภาษาจีน งานวิจัยทำการสร้างฐานความรู้ให้กับเครื่องโดยกำหนด Start Tag ให้กับตัวอักษรตัวแรก และกำหนด Non-start Tag ให้กับตัวอักษรตัวถัดไปในคำ นอกจากนี้มีการกำหนดหน้าที่ของคำให้กับแต่ละตัวอักษรด้วย จากผลงานวิจัยพบว่า CRFs สามารถช่วยเพิ่มประสิทธิภาพในการตัดคำในภาษาจีน ได้ค่า F1 สูงถึง 95%

3.งานวิจัยของ ชูชาติ หฤไชยะศักดิ์, ศราวุธ คงยัง และชัยอนันต์ ดำรงรัตน์ (2008) (พัชรินทร์ ทองอารีย์, 2548,น.23)

จากงานวิจัย สรุปออกมาว่าวิธี Conditional Random Field เป็นวิธีตัดคำแบบใช้เครื่องเรียนรู้ที่ดีที่สุด โดยงานวิจัยนี้ก็จะมุ่งประเด็นไปที่การรวมวิธีการตัดคำแบบ Conditional Random Field กับ การตัดคำแบบใช้พจนานุกรม เพื่อที่จะแก้ไขปัญหาในส่วนของการตัดคำที่กำกวม โดยเรียกวิธีนี้ว่า LearnLexTo

ภาพที่ 2.5

ภาพแสดงตัวอย่างการวิจัยของ LearnLexto



LearnLexTo จะให้ CRFs เป็นขั้นตอนหลักในการตัดคำ และให้วิธีที่ใช้พจนานุกรมเป็นตัวตรวจสอบข้อมูลที่ถูกตัดมาจาก CRFs โดยแสดงรายละเอียดดังภาพที่ 2.4 ในขั้นตอนแรกของการทดลองจะมีการประมาณค่าประสิทธิภาพของ CRFs ถึงความเหมาะสมในการใช้จำนวนแกรม (Gram) จากการทดลองสรุปออกมาว่าค่า 11-Gram มีความเหมาะสมมากที่สุด และได้ค่า F-measure 85.7% ขั้นตอนถัดมาคือ การทดลองประสิทธิภาพของ LearnLexTo ซึ่งผลที่ได้คือ LearnLexTo มีค่า F-measure 87.67% ในขณะที่การตัดคำแบบใช้พจนานุกรมอย่างเดียวมีค่า F-measure เพียง 82.71%

4.งานวิจัยของ ชูชาติ หฤไชยะศักดิ์, ศราวุธ คงยัง และ Matthew N. Daily (2008) (พัชรินทร์ ทองอารีย์,2548,น.23)

งานวิจัยนี้เป็นงานวิจัยที่ทำการศึกษาและเปรียบเทียบวิธีการตัดคำวิธีต่างๆ ในภาษาไทย โดยแยกประเภทของวิธีการตัดคำออกเป็น 2 แบบคือ แบบใช้พจนานุกรม (Dictionary-Based) และแบบใช้เครื่องเรียนรู้ (Machine Learning-Based)

แบบที่ใช้พจนานุกรม ก็เลือกใช้เทคนิคการตัดคำแบบเลือกคำยาวที่สุด และการตัดคำ แบบเลือกคำสอดคล้อง ส่วนแบบใช้เครื่องเรียนรู้จะเลือกใช้วิธี Naive Bayes (NB), Decision Tree และ Conditional Random Fields แต่อย่างไรก็ตามวิธีตัดคำที่มีประสิทธิภาพสูงที่สุดในงานวิจัยนี้คือวิธี Conditional Random Fields เพราะมีค่า F-measure 95.38%

แต่อย่างไรก็ตามการตัดคำด้วย วิธี Conditional Random Fields อาจจะเป็นวิธีที่ดีที่สุดในการตัดคำในตอนนี้แต่ก็ยังมีปัญหาการตัดคำที่เป็นนิพจน์ระบุนาม (Named Entity) ดังตัวอย่างภาพที่ 2.5

ภาพที่ 2.6

ตัวอย่างการตัดคำด้วย CRFs

เมือง|เธสซาโลนิกี| |กลุ่ม|คน|ร้าย|ขว้าง|ถัง|แก๊ส|ขนาด|เล็ก|และ|น้ำมัน| |ส่ง|ผล|ให้|เกิด|แรง|ระเบิด| สร้าง|ความ|เสียหาย|เล็กน้อย| |แต่|ขณะ|เกิด|เหตุ|ไม่|มี|คน|อยู่|ใน|สำนักงาน|จึง|ไม่|มี|ผู้|ได้|รับ| บาดเจ็บ| ขณะ|ที่|รัฐบาล|กรีซ|ซึ่ง|มี|เสียง|ข้าง|มาก|เกิน|เพียง| |1| |เสียง|กำลัง|อยู่|ใน|สภาพ| ง่อนแง่น|หลัง|จาก|เกิด|เหตุ|จลาจล|ร้าย|แรง|ที่สุด|ใน|รอบ|ทศวรรษ| |หลาย|ครั้ง|เมื่อ|เดือน| |ธ.ค.|ปี| ก่อน| |โดย|มี|ชนวน|เหตุ|จาก|การ|ที่|ตำรวจ|ยิง|วัย|รุ่น|เสีย|ชีวิต| |ผสม|กับ|กระแส|ไม่|พอใจ|เรื่อง|วัย| รุ่น|ว่าง|งาน|และ|มาตรการ|ทาง|เศรษฐกิจ|ของ|รัฐบาล| ล่าสุด|เมื่อ| |2| |สัปดาห์| |ก่อน|สำนัก|งาน| ของ|พรรค|รัฐบาล|ใน|กรุง|เอเธนส์| |ถูก|ขว้าง|ระเบิด|เสียหาย|เล็กน้อย|

|ความ|เข้มแข็ง| |อย่าง|เช่น| |จอมพล| |ป| |.| |พิบูลสงคราม| |และ| |พรรคพวก| |(|รวม|ทั้ง|นาย| ปรีดี| |พนมยงค์|ด้วย|)| |พลัง|ของ|ความ|คิด|เรื่อง|"|ชาติ|"| |ใน|ระยะ|ก่อน|สงคราม|โลก|ครั้ง|ที่| |2| |มี| ผล|ต่อ|พวก|เขา| |ทำ|ให้|จอมพล| |ป.| |และ|พรรคพวก|

คำที่ขีดเส้นได้เป็นนิพจน์ระบุนาม เป็นการตัดคำที่ออกมาแล้วไม่ถูกต้อง ซึ่งคำตอบที่สามารถตัดคำได้ถูกต้อง เช่น |เมืองเธสซาโลนิกี| และ |กรุงเอเธนส์| |จอมพล ป. พิบูลสงคราม| หรือ |นายปรีดี พนมยงค์| เป็นต้น

2.4.3.3 งานวิจัยที่เกี่ยวกับการจดจำนิพจน์ระบุนามภาษาไทย

1.งานวิจัยของ อัศนีย์ ก่อตระกูล และหัชทัย ชาญเลขา 2547:

การสกัดนิพจน์ระบุนามในภาษาไทย โดยใช้แบบจำลองทางสถิติร่วมกับฐานความรู้ เป็นการใช้หลักการผสม ระหว่างการใช้การคำนวณเชิงสถิติจากการฝึกฝนระบบ (Machine Learning) ร่วม กับการใช้ฐานความรู้ ซึ่งได้แก่ กฎฮิวริสติก และพจนานุกรมชื่อ ช่วยเพิ่มประสิทธิภาพของการ สกัดนิพจน์ระบุนาม (Named Entity) ประเภทชื่อบุคคล ชื่อองค์กร และชื่อสถานที่ ให้ค่า F เป็น 91.61% 88.53% และ 83.17%

อย่างไรก็ดี การสกัด นิพจน์ระบุนามวิธีนี้ ใช้คลัง พจนานุกรมชื่อ ขนาด 36084 คำ ทำให้สามารถในการสกัดนิพจน์ระบุนาม (Named Entity) ได้ดี

2.งานวิจัยของ วิโรจน์ อรุณมานะกุลและสุฤดี ฉัตรไตรมงคล 2548

งานวิจัย การรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยจะใช้วิธีการทางสถิติโดยใช้ N-Gram โดยรวมกับกฎ หลักการร่วมกันคือการหาค่าความสัมพันธ์ที่มีระหว่างหน่วยย่อย หากมีค่ามากโอกาสที่จะเป็นหน่วยเดียวกันก็จะมีสูงตามไปด้วย โดยใช้คลังข้อความ มีชื่อเฉพาะที่สำคัญหลักๆ ในบทความครบทั้ง 3 ประเภท ได้แก่ ชื่อคน ชื่อ สถานที่และชื่อองค์กร วัตถุประสงค์ของงานวิจัยนี้ทั่วไป โดยการเก็บรวบรวมจะเก็บ บทความข่าวจนกว่าจะได้ชื่อคน ชื่อองค์กร และชื่อสถานที่เกิน 3,000 ชื่อขึ้นไป และจะมีการแยกพยางค์ของคำแต่ละคำออกมา ในส่วนของกฎ

2.1 หลักฐานภายในที่ปรากฏอยู่หน้าชื่อเฉพาะ ได้แก่คำนำหน้าชื่อเฉพาะต่างๆ ซึ่งรายการคำนำหน้าชื่อที่เก็บได้มาจากคลังข้อมูล เช่น "อ." , "นาย" , "นาง" , "ดร." , "พล.ต." , "พล.ท." , "พ.ญ." , "พ.อ." , "ส.ก." , "ร.ต." , "ร.อ." , "ภ.ญ." , "ร.ท." , "แม่อุ้ย" , "จอมพล" , "น.ส." , "นางสาว" , "น.ต." , "อาจารย์" , "หัวหน้า" ,"พล.อ." , "น.พ." , "ม.ล." , "คุณหญิง" ,"หลวงพ่อ" , "ร.อ.หญิง" , "พ.ต.ท." , "พ.ต.หญิง" ,"ส.ต.อ." , "พล.ร.ท." , "ร.ต.อ." , "พล.ต.ท." , "ม.ร.ว." , "พล.อ.อ." , "พล.ต.ต." , "พล.ต.อ." ,"ส.ต.ต." , "พ.ต.ต." , "พ.ต.อ." , "พล.ต.ท." , "ร.ต.ต" , "จ.ส.ต." , "ร.ต.ท." , "พ.ญ.คุณหญิง" ,"ว่าที่ร.ต." , "พ.ต.อ.น.พ." , "กรม" , "พรรค" , "องค์การ" , "กระทรวง" , "สำนักงาน" ," สำนัก" ,"กองทัพ" , "บริษัท" , "ธนาคาร" , "องค์กร" , "กอง" , "โรงเรียน" , "โรงพยาบาล" ,"วิทยาลัย" , "บรรษัท" , "สหภาพ" , "สภา" , "มูลนิธิ" , "ศาล" , "สถาบัน" , "ศูนย์" , "ทวีป" ,"ร้านกาแฟ" , "ร้านอาหาร" , "สวนสัตว์" , "ค่าย" , "สถานทูต" ,"สนามกอล์ฟ" , "ท่าอากาศยาน" , "สุเหร่า" , "บึง" , "คุก" , "มัสยิด" , "สวนสาธารณะ" , "ร.ร." , "พระธาตุ" , "เกาะ" ,"สโมสร" , "สมาคม" , "สาธารณรัฐ" , "ราชอาณาจักร" , "จังหวัด" , "แม่น้ำ" , "ลุ่มน้ำ" ,"เขา" , "อ่างเก็บน้ำ" , "ท่าน้ำ" , "ท่าเรือ" , "ท่าเรือน้ำลึก" , "ปราสาท" , "ปราสาทหิน" ,"คลอง" , "วัด" , "วัง" , "รพ." , "อนุสาวรีย์" , "พระบรมราชานุสาวรีย์" , "เขื่อน" , "อนุสรณ์" ,"มณฑล" , "นคร" , "ต." , "อาคาร" , "ตึก" , "โรงแรม" , "สนามบิน" , "ประเทศ" , "แขวง" ,"หมู่บ้าน" , "สนามม้า" , "แฟลต" , "ซอย" , "ร้าน" , "เคหะชุมชน" , "โรงงาน" , "สวน" ,"บ้าน" , "ธนาคาร" , "ศาลเจ้า" , "สน." , "อุทยาน" , "ถ." , "เขตเทศบาลตำบล" , "ศูนย์ประชุม" , "หอประชุมอำเภอ" , "ทางด่วนสาย" , "ถนนสาย" , "สำนักงานสาขา" ,"หอประชุม" , "ที่ว่าการ" , "มูลนิธิ" , "ภูมิภาค" , "สถานี" , "ถนน" , "ที่ทำการสาขา" , "ห้อง" ,"สถาบัน" , "อำเภอ" , "หอ" , "ที่ทำการ" , "ภาค" , "ตลาดสดเทศบาล" , "ตลาดสด" ,"ตลาด" , "สนามกีฬาเทศบาล" , "สนามกีฬากลาง" , "สนามกีฬา" , "สนาม" , "ศาลา" ,"จ." , "เมือง" , "กรุง" , "เขต"

2.2 หลักฐานภายในที่ปรากฏอยู่ด้านหน้าและหลังชื่อเฉพาะ ได้แก่ บริษัท.....จำกัด

2.3 หลักฐานภายในที่อาจปรากฏอยู่ด้านหน้าหรือด้านหลังชื่อเฉพาะก็ได้ ได้แก่ มหาวิทยาลัย เช่น มหาวิทยาลัยธรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

จะเห็นได้ว่า ระบบที่ใช้วิธีทางสถิตินั้น ในงานนี้มีการใช้ คลังข้อมูลที่ผ่านการตัดพยางค์ ซึ่งทำให้ชื่อเฉพาะที่เลือกจากกลุ่มพยางค์ด้วยวิธีการทางสถิติที่ได้ มีจำนวนมาก ในการรู้จำชื่อเฉพาะนั้น มีหลักการร่วมกันคือการหาค่าความสัมพันธ์ที่มีระหว่างหน่วยย่อย หากมีค่ามากโอกาสที่จะเป็น หน่วยเดียวกันก็จะมีสูงตามไปด้วยเช่นคำว่า “สะดวก” ในคลังข้อมูลที่จะใช้ในวิธีทางสถิติ คำนี้จะถูกตัดพยางค์ออกเป็น “สะ-ดวก” ทำให้ได้พยางค์ 2 พยางค์ซึ่งพบว่ามีค่าความสัมพันธ์ระหว่างหน่วยย่อยทั้งสองมีมาก ทำให้ในจำนวนของคำตอบ (response) ที่วิธีการทางสถิติคัดเลือกออกมา จึงมีคำว่า “สะดวก” ติดมาด้วย ซึ่งคำดังกล่าวก็มิได้เป็นชื่อเฉพาะแต่อย่างใด จึง ทำให้ค่าความแม่นยำที่คำนวณจึงมีค่าต่ำ เพราะวิธีทางสถิติเหล่านี้ก็ยังมีข้อจำกัดประการหนึ่งคือ กรณีที่ชื่อ เฉพาะเป็นคำที่มีเพียงพยางค์เดียว เช่น <person>ชวน</person> หรือ <location> ไทย</location> จะไม่สามารถผ่านเงื่อนไขของวิธีทางสถิติที่พยายามหาค่าความสัมพันธ์ระหว่างพยางค์ เพราะชื่อเฉพาะที่จะผ่านเงื่อนไขไปได้ต้องมีไม่ต่ำกว่า 2 พยางค์ขึ้นไป โดยชื่อเฉพาะที่มี เพียงพยางค์เดียวจะมีจำนวน 1028 จากชื่อเฉพาะจำนวน 11683 ชื่อ โดยคิดเป็น 8.799% ของชื่อเฉพาะทั้งหมด ผลที่ได้ การรู้จำ ค่า F1 สำหรับชื่อเฉพาะประเภทชื่อคน 69.15% ชื่อองค์กร 62.95% และชื่อสถานที่ 38.87% ตามลำดับ

ตารางที่ 2.1 แสดงการเปรียบเทียบข้อดี – ข้อเสีย ของเทคนิคต่างๆ ที่ใช้ในการตัดคำงานวิจัยที่เกี่ยวกับการรู้จำชื่อเฉพาะภาษาไทย

ตารางที่ 2.3

ตารางการเปรียบเทียบเทคนิคการวิจัย

งานวิจัย	เทคนิคที่ใช้	ข้อดี	ข้อเสีย
สุรินทร์ จรรยาพรพงษ์	การใช้กฎในการตัดคำ กฎแบ่งได้เป็น 2 ชนิด คือ กฎการหาขอบเขตหน้า และกฎการหาขอบเขตหลัง	สามารถตัดคำได้โดยการใช้กฎ มีความรวดเร็วในการตัดคำ	สามารถตัดคำให้อยู่ในรูปแบบพยางค์หรือคำสั้นๆ เพียงแค่นั้น

(ต่อ)

งานวิจัย	เทคนิคที่ใช้	ข้อดี	ข้อเสีย
ยืน ภู่วรรณ และ วิวรรธ์ อิ่มอารมณ์ (2529)	การตัดคำแบบเลือกคำยาวที่สุด (Longest Matching)	สามารถย้อนกลับไปเพื่อเลือกคำอีกครั้งได้	เมื่อย้อนกลับไปหาคำนั้นจากพจนานุกรมแล้วยังไม่พบก็จะทำให้เสียเวลา
กานดา รุณนะพงศา และ ปโยธร อุราธรรมกุล(2549)	การใช้กฎและพจนานุกรมแบบใหม่	พจนานุกรมแบบใหม่สามารถ ตัดคำได้ดีมากขึ้น	ยังไม่สามาถตัดคำเฉพาะหรือคำที่มีความยาวหลายพยางค์
อัศนีย์ ก่อตระกูลและคณะ (2538)	Tri – Gram	สามารถลดรูปแบบการตัดคำที่ไม่เหมาะสม	ต้องใช้คลังข้อความขนาดใหญ่มาก
John Lafferty, Andrew McCallum และ Fernando Pereira (2001)	CRFs	นำ CRFs เข้ามาประยุกต์ใช้กับการแบ่งข้อมูลที่อยู่ติดกันและได้ผลดี	ใช้กับข้อมูลที่เป็นภาษาอังกฤษ
Fuchun Peng, Fangfang Feng และ Andrew McCallum (2004)	CRFs	นำ CRFs ไปประยุกต์ใช้กับการตัดคำของภาษาจีนและช่วยเพิ่มความถูกต้องในการตัดคำ	ใช้กับข้อมูลที่เป็นภาษาจีน
ชูชาติ หฤไชยะศักดิ์, ศราวุธ คงยังและชัยอนันต์ ดำรงรัตน์ (2008)	LearnLexTo คือการตัดคำด้วย CRFs แต่ตรวจสอบด้วยพจนานุกรม	เพิ่มประสิทธิภาพในการตัดคำ	การใช้พจนานุกรมเข้ามาตรวจสอบสามารถเพิ่มประสิทธิภาพขึ้นเล็กน้อยจากการตัดคำด้วย CRFs

(ต่อ)

งานวิจัย	เทคนิคที่ใช้	ข้อดี	ข้อเสีย
ชูชาติ หฤไชยะศักดิ์ ,ศราวุธ คงยัง และ Matthew N.Daily	DT,NB,SVM และ CRF	เปรียบเทียบให้เห็นว่าวิธีการตัดคำวิธีไหนมีความถูกต้องมากที่สุด	เปรียบเทียบเฉพาะการตัดคำเท่านั้น
อัศนีย์ ก่อตระกูล และ หัชทัย ชาญเลขา (2547)	โดยใช้ฮิวริสติกร่วมกับฐานความรู้และข้อมูลเชิงสถิติและใช้พจนานุกรมชื่อ	ช่วยเพิ่มประสิทธิภาพของการ สกัดนิพจน์ระบุนาม (Named Entity)	ต้องมีพจนานุกรมชื่อจำนวนเยอะ
วิโรจน์ อรุณมานะกุล และสุฤดี ฉัตรไตรมงคล (2548)	N-Gram โดยรวมกับกฎ	สามารถสกัดนิพจน์ระบุนามได้	วิธีทางสถิติก็ยังมีข้อจำกัดประการหนึ่งคือ กรณีที่นิพจน์ระบุนาม เป็นคำที่ มีเพียงพยางค์เดียว

จากตารางที่ 2.1 จะเห็นได้ว่าทั้งเทคนิคการตัดคำและงานวิจัยที่เกี่ยวกับการการจดจำนิพจน์ระบุนาม มีผู้วิจัยและพัฒนาวิธีการต่างๆ มาอย่างต่อเนื่อง และมีการนำเสนอว่าแต่ละเทคนิคนั้นมีข้อดีและข้อเสียอย่างไร เทคนิคไหนมีประสิทธิภาพมากที่สุด โดยจากงานวิจัยต่าง ๆ ที่ผ่านมาสามารถสรุปได้ว่า

วิธีการตัดคำโดยใช้การเรียนรู้ของเครื่องสามารถตัดคำได้มีประสิทธิภาพกว่าการตัดคำแบบใช้พจนานุกรม และการตัดคำโดยใช้การเรียนรู้ของเครื่องด้วยเทคนิค CRFs นั้นมีประสิทธิภาพดีที่สุดในทั้งภาษาไทย ภาษาอังกฤษ และภาษาจีน และจะเห็นได้ว่านิพจน์ระบุนามในภาษาไทย เป็นประเด็นที่ยังเป็นปัญหาและมีความสำคัญต่อการประมวลผลภาษาในการพัฒนางานด้านต่างๆ จึงต้องมีการพัฒนาระบบประมวลผลให้มีประสิทธิภาพทั้งในด้านความถูกต้องด้านภาษาและด้านความแม่นยำรวดเร็ว ปัญหาที่กล่าวถึงนั้นเนื่องมาจากโดยธรรมชาติของนิพจน์ระบุนามแล้วมีแนวโน้มที่จะเกิดขึ้นใหม่อยู่ตลอดเวลา การใช้วิธีทำให้คอมพิวเตอร์รู้จักนิพจน์ระบุนามโดยการใช้คลังคำศัพท์หรือพจนานุกรมจึงไม่สามารถครอบคลุมนิพจน์ระบุนามที่เกิดขึ้นใหม่

นิพจน์ระบุนามเองที่มีรูปแบบในลักษณะเดียวกับคำนามทั่วไป เป็นผลให้ระบบคอมพิวเตอร์ไม่สามารถแยกแยะนิพจน์ระบุนามออกจากคำนามทั่วไปได้ ผลจากการที่คอมพิวเตอร์ไม่รู้จักและไม่ สามารถแยก แยะนิพจน์ระบุนามได้ ทำให้ความสามารถในการประมวลผลภาษาโดยรวมของเครื่องคอมพิวเตอร์มีประสิทธิภาพลง การศึกษาหาวิธีการที่จะทำให้คอมพิวเตอร์สามารถแยกแยะและบ่งชี้ได้ว่าหน่วยภาษานั้นเป็นนิพจน์ระบุนามได้จึงเป็นเรื่องสำคัญเรื่องหนึ่งในการประมวลผลภาษาไทย